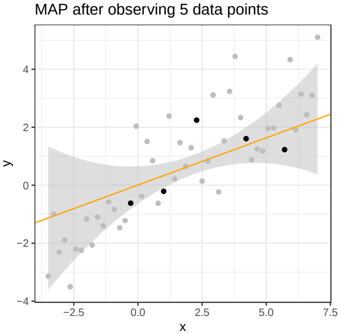
Introduction to Machine Learning

The Bayesian Linear Model

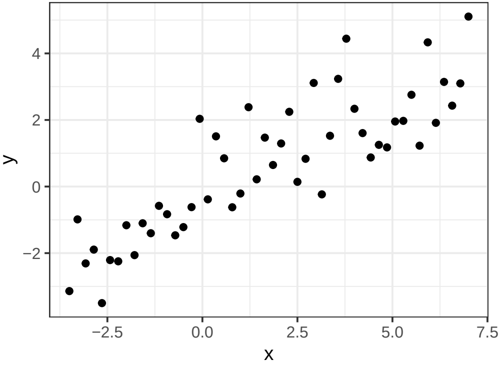


Learning goals

- Know the Bayesian linear model
- The Bayesian LM returns a (posterior) distribution instead of a point estimate
- Know how to derive the posterior distribution for a Bayesian LM

REVIEW: THE BAYESIAN LINEAR MODEL

Let $\mathcal{D} = \left\{(\mathbf{x}^{(1)}, y^{(1)}), ..., (\mathbf{x}^{(n)}, y^{(n)})\right\}$ be a training set of i.i.d. observations from some unknown distribution.



Let $\mathbf{y} = (y^{(1)}, ..., y^{(n)})^\top$ and $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the design matrix where the i-th row contains vector $\mathbf{x}^{(i)}$.

REVIEW: THE BAYESIAN LINEAR MODEL

The linear regression model is defined as

$$y = f(\mathbf{x}) + \epsilon = \boldsymbol{\theta}^T \mathbf{x} + \epsilon$$

or on the data:

$$y^{(i)} \quad = \quad f\left(\mathbf{x}^{(i)}\right) + \epsilon^{(i)} = \boldsymbol{\theta}^T \mathbf{x}^{(i)} + \epsilon^{(i)}, \quad \text{for } i \in \{1, \dots, n\}$$

We now assume (from a Bayesian perspective) that also our parameter vector $\boldsymbol{\theta}$ is stochastic and follows a distribution. The observed values $y^{(i)}$ differ from the function values $f\left(\mathbf{x}^{(i)}\right)$ by some additive noise, which is assumed to be i.i.d. Gaussian

$$\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$$

and independent of $\mathbf{x}$ and $\boldsymbol{\theta}$.

REVIEW: THE BAYESIAN LINEAR MODEL

Let us assume we have **prior beliefs** about the parameter $\boldsymbol{\theta}$ that are represented in a prior distribution $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{I}_p)$.

Whenever data points are observed, we update the parameters' prior distribution according to Bayes' rule

$$\underbrace{p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})}^{\text{likelihood}} \overbrace{q(\boldsymbol{\theta})}^{\text{prior}}}{\underbrace{p(\mathbf{y}|\mathbf{X})}_{\text{marginal}}}.$$

REVIEW: THE BAYESIAN LINEAR MODEL

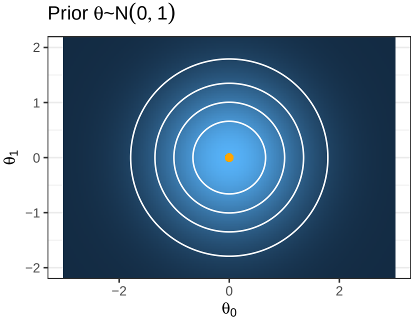
The posterior distribution of the parameter $\boldsymbol{\theta}$ is again normal distributed (the Gaussian family is self-conjugate):

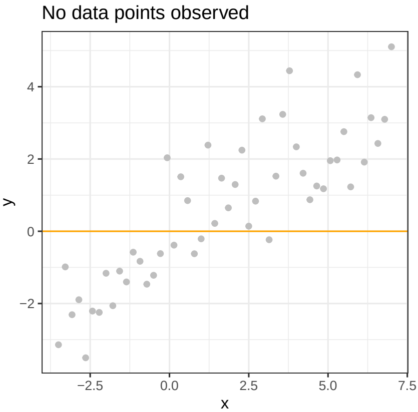
$$\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2}\boldsymbol{A}^{-1}\mathbf{X}^\top\mathbf{y}, \boldsymbol{A}^{-1})$$

with $\boldsymbol{A} := \sigma^{-2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\tau^2}\boldsymbol{I}_p$.

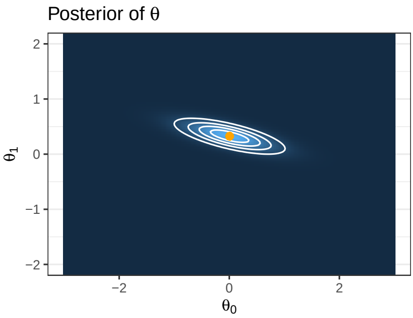
Note: If the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y})$ are in the same probability distribution family as the prior $q(\boldsymbol{\theta})$ w.r.t. a specific likelihood function $p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta})$, they are called **conjugate distributions**. The prior is then called a **conjugate prior** for the likelihood. The Gaussian family is self-conjugate: Choosing a Gaussian prior for a Gaussian Likelihood ensures that the posterior is Gaussian.

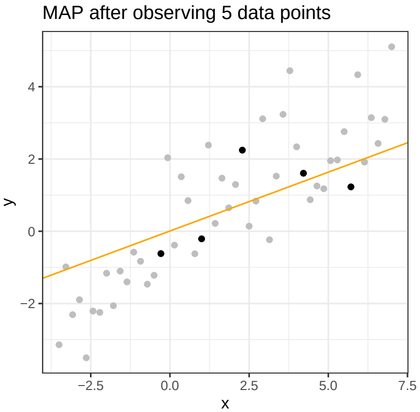
REVIEW: THE BAYESIAN LINEAR MODEL



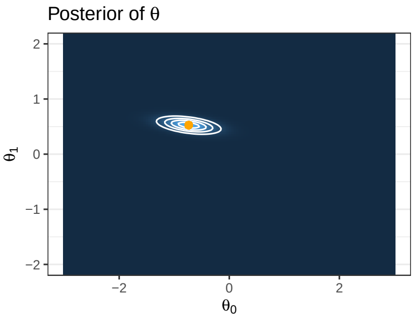


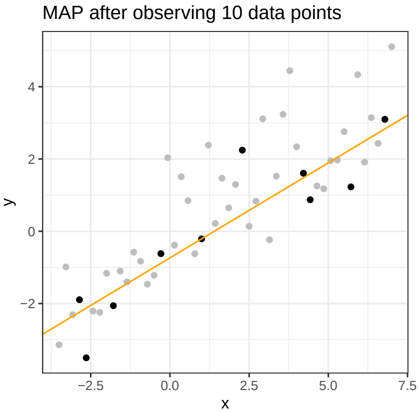
REVIEW: THE BAYESIAN LINEAR MODEL



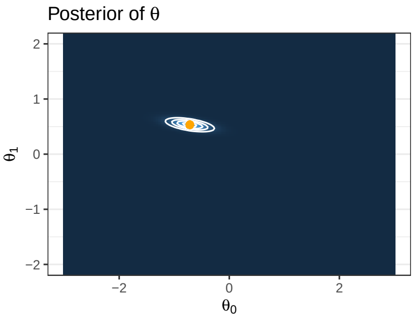


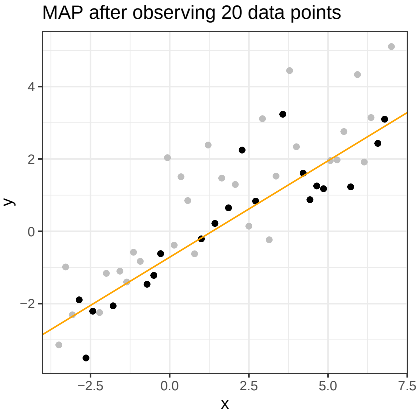
REVIEW: THE BAYESIAN LINEAR MODEL





REVIEW: THE BAYESIAN LINEAR MODEL





REVIEW: THE BAYESIAN LINEAR MODEL

Proof:
We want to show that

- for a Gaussian prior on $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{I}_p)$
- for a Gaussian Likelihood $y \mid \mathbf{X}, \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{X}^\top \boldsymbol{\theta}, \sigma^2 \boldsymbol{I}_n)$

the resulting posterior is Gaussian $\mathcal{N}(\sigma^{-2}\boldsymbol{A}^{-1}\mathbf{X}^\top \mathbf{y}, \boldsymbol{A}^{-1})$ with $\boldsymbol{A} := \sigma^{-2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\tau^2}\boldsymbol{I}_p$.
Plugging in Bayes' rule and multiplying out yields

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})q(\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}) - \frac{1}{2\tau^2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right] \\
&= \exp\left[-\frac{1}{2}\left(\underbrace{\sigma^{-2}\mathbf{y}^\top\mathbf{y}}_{\text{doesn't depend on } \boldsymbol{\theta}} - 2\sigma^{-2}\mathbf{y}^\top\mathbf{X}\boldsymbol{\theta} + \sigma^{-2}\boldsymbol{\theta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta} + \tau^{-2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\sigma^{-2}\boldsymbol{\theta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta} + \tau^{-2}\boldsymbol{\theta}^\top\boldsymbol{\theta} - 2\sigma^{-2}\mathbf{y}^\top\mathbf{X}\boldsymbol{\theta}\right)\right] \\
&= \exp\left[-\frac{1}{2}\boldsymbol{\theta}^\top\underbrace{\left(\sigma^{-2}\mathbf{X}^\top\mathbf{X} + \tau^{-2}\boldsymbol{I}_p\right)}_{:=\mathbf{A}}\boldsymbol{\theta} + \sigma^{-2}\mathbf{y}^\top\mathbf{X}\boldsymbol{\theta}\right]
\end{aligned}
$$

This expression resembles a normal density - except for the term in red!

REVIEW: THE BAYESIAN LINEAR MODEL

Note: We need not worry about the normalizing constant since its mere role is to convert probability functions to density functions with a total probability of one. We subtract a (not yet defined) constant c while compensating for this change by adding the respective terms ("adding 0"), emphasized in green:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) \quad &\propto \quad \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - c)^\top \mathbf{A}(\boldsymbol{\theta} - c) - c^\top \mathbf{A}\boldsymbol{\theta} + \underbrace{\frac{1}{2}c^\top \mathbf{A}c}_{\text{doesn't depend on } \boldsymbol{\theta}} + \sigma^{-2}\mathbf{y}^\top \mathbf{X}\boldsymbol{\theta}\right] \\
&\propto \quad \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - c)^\top \mathbf{A}(\boldsymbol{\theta} - c) - c^\top \mathbf{A}\boldsymbol{\theta} + \sigma^{-2}\mathbf{y}^\top \mathbf{X}\boldsymbol{\theta}\right]
\end{aligned}
$$

If we choose c such that $-c^\top \mathbf{A}\boldsymbol{\theta} + \sigma^{-2}\mathbf{y}^\top \mathbf{X}\boldsymbol{\theta} = 0$, the posterior is normal with mean c and covariance matrix $\mathbf{A}^{-1}$. Taking into account that $\mathbf{A}$ is symmetric, this is if we choose

$$
\begin{aligned}
& \sigma^{-2}\mathbf{y}^\top \mathbf{X} = c^\top \mathbf{A} \\
\Leftrightarrow \quad & \sigma^{-2}\mathbf{y}^\top \mathbf{X}\mathbf{A}^{-1} = c^\top \\
\Leftrightarrow \quad & c = \sigma^{-2}\mathbf{A}^{-1}\mathbf{X}^\top \mathbf{y}
\end{aligned}
$$

as claimed.

REVIEW: THE BAYESIAN LINEAR MODEL

Based on the posterior distribution

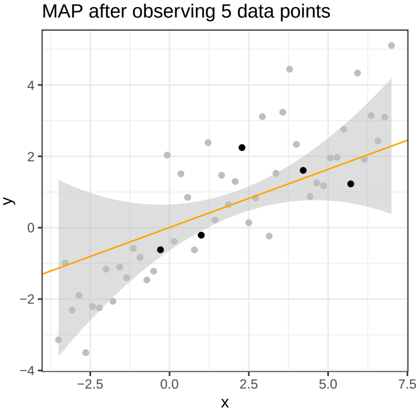
$$\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2}\boldsymbol{A}^{-1}\mathbf{X}^\top\mathbf{y}, \boldsymbol{A}^{-1})$$

we can derive the predictive distribution for a new observations $\mathbf{x}_*$. The predictive distribution for the Bayesian linear model, i.e. the distribution of $\boldsymbol{\theta}^\top\mathbf{x}_*$, is

$$y_* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_* \sim \mathcal{N}(\sigma^{-2}\mathbf{y}^\top\mathbf{X}\mathbf{A}^{-1}\mathbf{x}_*, \mathbf{x}_*^\top\mathbf{A}^{-1}\mathbf{x}_*)$$

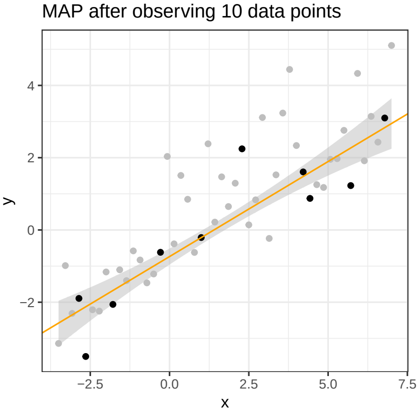
(applying the rules for linear transformations of Gaussians).

REVIEW: THE BAYESIAN LINEAR MODEL



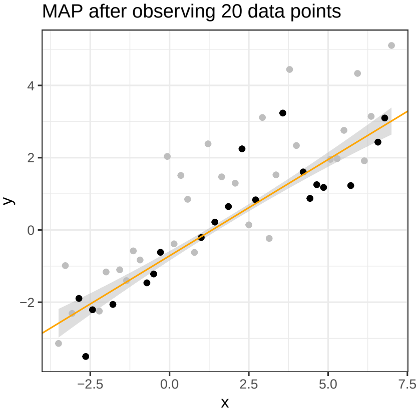
For every test input $\mathbf{x}_*$, we get a distribution over the prediction y_*. In particular, we get a posterior mean (orange) and a posterior variance (grey region equals $+/-$ two times standard deviation).

REVIEW: THE BAYESIAN LINEAR MODEL



For every test input $\mathbf{x}_*$, we get a distribution over the prediction y_*. In particular, we get a posterior mean (orange) and a posterior variance (grey region equals $+/-$ two times standard deviation).

REVIEW: THE BAYESIAN LINEAR MODEL



For every test input $\mathbf{x}_*$, we get a distribution over the prediction y_*. In particular, we get a posterior mean (orange) and a posterior variance (grey region equals $+/-$ two times standard deviation).

SUMMARY: THE BAYESIAN LINEAR MODEL

- By switching to a Bayesian perspective, we do not only have point estimates for the parameter $\boldsymbol{\theta}$, but whole **distributions**
- From the posterior distribution of $\boldsymbol{\theta}$, we can derive a predictive distribution for $y_* = \boldsymbol{\theta}^\top \mathbf{x}_*$.
- We can perform online updates: Whenever datapoints are observed, we can update the **posterior distribution** of $\boldsymbol{\theta}$

Next, we want to develop a theory for general shape functions, and not only for linear function.